Running Cufflinks with your own transcript annotation data

Panagiotis Papastamoulis

University of Manchester

The following example first uses `tophat` to align a set of Drosophila reads to the reference annotation. We then run `cufflinks` in order to estimate transcript expression. Finally, `cuffdiff` is used in order to detect differentially expressed transcripts between two conditions. The following modules should be loaded:

```
module load apps/binapps/tophat/2.0.9
module load apps/gcc/samtools/0.1.18
module load apps/binapps/cufflinks/2.1.1
module load apps/binapps/bowtie2/2.1.0
```

0.1 Reference annotation

When mapping with `tophat` against a reference annotation it is essential to download the reference genome by the Cufflinks annotation page. Any other user-defined annotation will almost certainly fail. This example deals with drosophila melanogaster, so the reference annotation is downloaded from

```
ftp://igenome:G3nom3s4u@ussd-ftp.illumina.com/Drosophila_melanogaster/
Ensembl/BDGP5/Drosophila_melanogaster_Ensembl_BDGP5.tar.gz
```

We extract all files from this link to a directory called `CufflinksAnnotationDirectory`.

0.2 Mapping the reads with `tophat`

Assume that our sample consists of two biological conditions (`A` and `B`) and each one consists of two sets of reads (`.fastq` files):
`conditionA_1.fastq conditionA_2.fastq`
`conditionB_1.fastq conditionB_2.fastq`

The following jobscript will map the four available samples to the reference annotation. No novel junctions will be examined.

```
## SGE Stuff
### Use the current/submission directory as the working directory
#$ -cwd
### Inherit the user environment settings from the login node
#$ -V
#$ -pe smp.pe 4 #### Run in parallel with 4 threads
export OMP_NUM_THREADS=$NSLOTS
```

```
# map sample conditionA_1.fastq allowing at most 35 multiple alignments
#    using the drosophila melanogaster reference genome.
#    All output is written to directory A1.tophat.out

tophat -o A1.tophat.out --no-novel-juncs -x 35 -T -p 4        \
-G CufflinksAnnotationDirectory/Drosophila_melanogaster/      \
Ensembl/BDGP5/Annotation/Genes/genes.gtf                      \
--transcriptome-index=transcriptome_data/known                \
CufflinksAnnotationDirectory/Drosophila_melanogaster/Ensembl/ \
BDGP5/Sequence/Bowtie2Index/genome conditionA_1.fastq

# Note that now the transcriptome_data directory is in the current
#    directory containing Bowtie index files. Then for subsequent TopHat
#    runs with the same genome and known transcripts but different reads
#    the -G option is no longer needed.

# do the same for the rest samples:
# sample  conditionA_2.fastq
tophat -o A2.tophat.out --no-novel-juncs -x 35 -T -p 4        \
--transcriptome-index=transcriptome_data/known                \
CufflinksAnnotationDirectory/Drosophila_melanogaster/Ensembl/ \
BDGP5/Sequence/Bowtie2Index/genome conditionA_2.fastq

# sample  conditionB_1.fastq
tophat -o B1.tophat.out --no-novel-juncs -x 35 -T -p 4        \
--transcriptome-index=transcriptome_data/known                \
CufflinksAnnotationDirectory/Drosophila_melanogaster/Ensembl/ \
BDGP5/Sequence/Bowtie2Index/genome conditionB_1.fastq

# sample  conditionB_2.fastq
tophat -o B2.tophat.out --no-novel-juncs -x 35 -T -p 4        \
--transcriptome-index=transcriptome_data/known                \
CufflinksAnnotationDirectory/Drosophila_melanogaster/Ensembl/ \
BDGP5/Sequence/Bowtie2Index/genome conditionB_2.fastq
```

Note that after the first run (for sample A1) the `transcriptome_data` directory is in the current directory containing Bowtie index files. Then for subsequent TopHat runs with the same genome and known transcripts but different reads (samples A1, B1 and B2) the `-G` option is no longer needed.

0.3 `cufflinks` and `cuffdiff`

Now we can estimate expression and perform DE analysis using `cufflinks` and `cuffdiff`.

```
## SGE Stuff
### Use the current/submission directory as the working directory
#$ -cwd
### Inherit the user environment settings from the login node
#$ -V
#$ -pe smp.pe 4 #### Run in parallel with 4 threads
```

```
export OMP_NUM_THREADS=$NSLOTS

#estimate expression for sample A1. Output written to A1.cufflinks.out
cufflinks -q -o A1.cufflinks.out -u -p 4                      \
-G CufflinksAnnotationDirectory/Drosophila_melanogaster/      \
Ensembl\BDGP5/Annotation/Genes/genes.gtf                      \
A1.tophat.out/accepted_hits.bam >> A1_cuff.log

#estimate expression for sample A2. Output written to A2.cufflinks.out
cufflinks -q -o A2.cufflinks.out -u -p 4                      \
-G CufflinksAnnotationDirectory/Drosophila_melanogaster/      \
Ensembl\BDGP5/Annotation/Genes/genes.gtf                      \
A2.tophat.out/accepted_hits.bam >> A2_cuff.log

#estimate expression for sample B1. Output written to B1.cufflinks.out
cufflinks -q -o B1.cufflinks.out -u -p 4                      \
-G CufflinksAnnotationDirectory/Drosophila_melanogaster/      \
Ensembl\BDGP5/Annotation/Genes/genes.gtf                      \
B1.tophat.out/accepted_hits.bam >> B1_cuff.log

#estimate expression for sample B2. Output written to B2.cufflinks.out
cufflinks -q -o B2.cufflinks.out -u -p 4                      \
-G CufflinksAnnotationDirectory/Drosophila_melanogaster/      \
Ensembl\BDGP5/Annotation/Genes/genes.gtf                      \
B2.tophat.out/accepted_hits.bam >> B2_cuff.log

# Perform Differential Expression analysis between the two conditions:

cuffcompare -o cuff_compare A1.cufflinks.out/transcripts.gtf      \
A2.cufflinks.out/transcripts.gtf B1.cufflinks.out/transcripts.gtf \
B2.cufflinks.out/transcripts.gtf

cuffdiff -o cuffdif_out -p 4 cuff_compare.combined.gtf            \
A1.tophat.out/accepted_hits.bam,A2.tophat.out/accepted_hits.bam   \
B1.tophat.out/accepted_hits.bam,B2.tophat.out/accepted_hits.bam
```

Note that the DE analysis results are saved to the `cuffdif_out` directory.